Deep Learning Classification and Quantification of Pejorative and Nonpejorative Architectures in Resected Hepatocellular Carcinoma from Digital Histopathologic Images

Astrid Laurent-Bellue, Aymen Sadraoui, Laura Claude, Julien Calderaro, Katia Posseme, Eric Vibert, Daniel Cherqui, Olivier Rosmorduc, Maïté Lewin, Jean-Christophe Pesquet, et al.

Title

Deep learning classification and quantification of pejorative and non-pejorative architectures in resected hepatocellular carcinoma from digital histopathological images

Short running head

Deep Learning for HCC

Authors

Astrid Laurent-Bellue*, MD

Department of Pathology, Bicêtre Hospital, Assistance Publique-Hôpitaux de Paris

78 rue du Général Leclerc, 94270 Le Kremlin-Bicêtre, France

Faculté de Médecine, Paris-Saclay University

63 rue Gabriel Péri, 94270, Le Kremlin-Bicêtre, France

UMR 1193, Paris-Saclay University, INSERM

12 avenue Paul Vaillant-Couturier, 94800, Villejuif, France

Aymen Sadraoui*

Centre de Vision Numérique, Paris-Saclay University, Inria, CentraleSupélec

3 rue Joliot-Curie, 91190, Gif-sur-Yvette, France

Laura Claude, MD

Department of Pathology, Charles Nicolle Hospital, Rouen

37 boulevard Gambetta, 76000, Rouen, France

Julien Calderaro, MD, PhD

Department of Pathology, Henri-Mondor Hospital, Assistance Publique-Hôpitaux de Paris

1 rue Gustave Eiffel, 94000, Créteil, France

Katia Posseme

Department of Pathology, Bicêtre Hospital, Assistance Publique-Hôpitaux de Paris

78 rue du Général Leclerc, 94270 Le Kremlin-Bicêtre, France

Eric Vibert, MD, PhD

Centre Hépato-Biliaire, Paul-Brousse Hospital, Assistance Publique-Hôpitaux de Paris

12 avenue Paul Vaillant-Couturier, 94800, Villejuif, France

Faculté de Médecine, Paris-Saclay University

63 rue Gabriel Péri, 94270, Le Kremlin-Bicêtre, France

UMR 1193, Paris-Saclay University, INSERM

12 avenue Paul Vaillant-Couturier, 94800, Villejuif, France

Daniel Cherqui, MD, PhD

Centre Hépato-Biliaire, Paul-Brousse Hospital, Assistance Publique-Hôpitaux de Paris

12 avenue Paul Vaillant-Couturier, 94800, Villejuif, France

Faculté de Médecine, Paris-Saclay University

63 rue Gabriel Péri, 94270, Le Kremlin-Bicêtre, France

UMR 1193, Paris-Saclay University, INSERM

12 avenue Paul Vaillant-Couturier, 94800, Villejuif, France

Olivier Rosmorduc, MD, PhD

Centre Hépato-Biliaire, Paul-Brousse Hospital, Assistance Publique-Hôpitaux de Paris

12 avenue Paul Vaillant-Couturier, 94800, Villejuif, France

UMR 1193, Paris-Saclay University, INSERM

12 avenue Paul Vaillant-Couturier, 94800, Villejuif, France

Maïté Lewin, MD, PhD

Department of Radiology, Paul-Brousse Hospital, Assistance Publique-Hôpitaux de Paris

12 avenue Paul Vaillant-Couturier, 94800, Villejuif, France

Faculté de Médecine, Paris-Saclay University

63 rue Gabriel Péri, 94270, Le Kremlin-Bicêtre, France

UMR 1193, Paris-Saclay University, INSERM

12 avenue Paul Vaillant-Couturier, 94800, Villejuif, France

Jean-Christophe Pesquet, PhD

Centre de Vision Numérique, Paris-Saclay University, Inria, CentraleSupélec

3 rue Joliot-Curie, 91190, Gif-sur-Yvette, France

Catherine Guettier, MD, PhD

Department of Pathology, Bicêtre Hospital, Assistance Publique-Hôpitaux de Paris

78 rue du Général Leclerc, 94270 Le Kremlin-Bicêtre, France

Faculté de Médecine, Paris-Saclay University

63 rue Gabriel Péri, 94270, Le Kremlin-Bicêtre, France

UMR 1193, Paris-Saclay University, INSERM

12 avenue Paul Vaillant-Couturier, 94800, Villejuif, France

*These authors contributed equally to this work

Corresponding author:

Catherine Guettier, MD, PhD

Department of Pathology, Bicêtre Hospital, Assistance Publique-Hôpitaux de Paris,

78 rue du Général Leclerc, 94270, Le Kremlin-Bicêtre, France

+33145212025

catherine.guettier@aphp.fr

Grant numbers and sources of support

The authors received no specific funding for this work.

Number of text pages: 43

Number of tables: 3

Number of figures: 7

Abstract (220 words)

Liver resection is one of the best treatments for small hepatocellular carcinoma, but post-resection recurrence is frequent. Biotherapies have emerged as an efficient adjuvant treatment making the identification of patients at high risk of recurrence critical. Microvascular invasion, poor differentiation, pejorative macrotrabecular, and “vessels encapsulating tumor clusters” architectures are the most accurate histological predictors of recurrence but their evaluation is time-consuming and imperfect. A supervised deep learning-based approach with ResNet34 on 680 Whole Slide Images from 107 liver resection specimens allowed to build an algorithm for the identification and quantification of these pejorative architectures. This model achieved an accuracy of 0.864 at patch-level and 0.823 at Whole Slide Image-level. To assess its robustness, it was validated on an external cohort of 29 hepatocellular carcinomas from another hospital with an accuracy of 0.787 at Whole Slide Image-level, affirming its generalization capabilities. Moreover, largest connected areas of the pejorative architectures extracted from the model were positively correlated to the presence of microvascular invasion and the number of tumor emboli. These results suggest that the identification of pejorative architectures could be an efficient surrogate of microvascular invasion and have a strong predictive value for the risk of recurrence. This study is the first step in the construction of a composite predictive algorithm for early post-resection recurrence of hepatocellular carcinoma, including artificial intelligence-based features.

List of nonstandard abbreviations

AI: artificial intelligence

HCC: hepatocellular carcinoma

HES: hematoxylin, eosin, and saffron

MTM: macrotrabecular massive

mVI: microvascular invasion

VETC: vessels encapsulating tumor clusters

WHO: World Health Organization

WSI: Whole Slide Image

Original article (6072 words)

Introduction

Hepatocellular carcinoma (HCC) is the main primary malignancy of the liver, ranking as the sixth most commonly diagnosed cancer and the third leading cause of cancer death worldwide in 2020, with approximately 906,000 new cases and 830,000 deaths[1]. In Western countries, HCC predominantly develops in the context of cirrhosis caused by non-alcoholic fatty liver disease, alcohol-related liver disease, or chronic viral hepatitis B or C. The prognosis for HCC remains poor with a 5-year survival rate of less than 10%[2]. Many cases are diagnosed at advanced stages making them inaccessible to curative treatment at the time of diagnosis. Even for resectable tumors, the rate of recurrence after surgical treatment exceeds 50% at 5 years post-liver resection[3,4]. Most recurrences occur early, within 2 years of resection, and are considered as true recurrences of the initial tumor and not as independent tumors[5].

Liver transplantation can be curative for both HCC and underlying cirrhosis with a lower recurrence rate of around 10%-15% at 5 years but the shortage of grafts limits access to this treatment and a large majority of patients are not eligible for transplantation. Chemotherapy is not efficient in HCC. Immunotherapy alone or in combination with targeted Tyrosine Kinase Inhibitors or antiangiogenic drugs emerged as a promising therapeutic strategy for advanced HCC[6,7]. The use of these biotherapies as post-resection adjuvant treatment is the next step in the therapeutic strategy for HCC. Interim analysis of a randomized multicentric phase 3 trial aiming to assess the efficacy of adjuvant atezolizumab plus bevacizumab versus active surveillance in patients with resected or ablated high-risk HCC demonstrated a 28% lower risk of recurrence or death in patients receiving adjuvant atezolizumab plus bevacizumab[8]. Thus, the identification of patients at high risk of relapse is critical for the design of adjuvant studies because post-operative therapy might particularly benefit this subgroup. Pre-operative risk stratification of patients eligible for liver resection is still suboptimal although pre-operating imaging demonstrated acceptable performance as potential surrogates of pejorative histological features, thanks to technological

progress and radiomic approaches[9,10]. Nevertheless, the current gold standard for the identification of the most accurate prognosis indicators is the pathological examination of surgical specimens.

Indeed, microvascular invasion (mVI)[11], poor histological differentiation, macrotrabecular architecture[12,13] or “vessels encapsulating tumor clusters” (VETC) architecture[14,15] have been identified as highly predictive of HCC recurrence risk. Identification of these histopronostic pejorative factors requires extensive sampling of the tumor and its adjacent parenchyma and careful examination of microscopic slides. Their evaluation of surgical specimens by conventional pathological examination is imperfect, time-consuming, and not very reproducible, in particular for the histological grade and the presence of mVI. Indeed, the broad histological heterogeneity of HCC makes the prognostic stratification of patients challenging. However, it has been shown that macrotrabecular and VETC architectures are correlated one with another and also with the presence of mVI[12,15]. These correlations have been demonstrated for a stringent definition of macrotrabecular massive (MTM) HCC subtype and VETC HCC phenotype, defined respectively by macrotrabecular growth pattern in >50% and VETC pattern in ≥55% of tumor, but the predictive value of less abundant pejorative components for the presence of mVI has not been assessed.

Recently, artificial intelligence (AI) applied to Whole Slide Images (WSIs) has emerged as a helpful, faster, and more reproducible approach to performing segmentation and classification of tumors including liver tumors[16–18].

This study aims to build a deep learning-based approach for automatic and robust identification and quantification of pejorative macrotrabecular and VETC HCC architecture on WSIs from liver surgical curative resection specimens and to assess the correlation between pejorative tumor architectures and the presence of mVI. It is the first step in the process of building a composite predictive model of the risk of post-resection recurrence of HCC, including AI-based features.

Materials and Methods

Ethics Approval

The study conforms to the General Data Protection Regulation and was approved by the Institutional Review Board of Mondor Hospital (IRB#00011558, notification number: 2022-135).

Patients

A series of 107 patients over 18 years old who underwent curative hepatic resection for HCC, excluding liver transplantation, between January 2016 and December 2020 in Paul-Brousse Hospital, France was retrospectively collected. None of these patients had received cancer treatment before surgery. Patients with R1 and R2 resection were excluded.

The following clinico-biological data were systematically collected for all patients: age, sex, etiology of the underlying liver disease, serum alpha-foetoprotein, albumin, and total bilirubin before surgery.

The following data (Table 1) were extracted from the pathological reports for all tumors: number of nodules and size, satellite nodules, multinodular expansive macroscopic pattern, histological grade, and HCC subtype according to the World Health Organization (WHO) Classification of Tumors: Digestive System Tumors 2019[19], presence of mVI, American Joint Committee on Cancer 8th edition stage[20] and METAVIR fibrosis score for non-tumor liver parenchyma[21].

An external validation cohort of 29 patients selected according to the same criteria from the Department of Pathology at Henri-Mondor Hospital, France, was used as an external validation for the model. The same clinico-biological and pathological data were systematically extracted from patient records.

Whole Slide Images

Surgical specimens were fixed with 4% neutral formaldehyde. For each tumor of the cohort, 2 to 21 blocks were sampled from tumoral and peri-tumoral tissue according to the tumor size with at least one block for 1 cm of the highest diameter. Blocks were embedded in paraffin and cut at 4 µm thick. Slides were stained with hematoxylin, eosin, and saffron (HES) in the Department of Pathology at Bicêtre Hospital.

All slides (n=680) were digitized at 20 × objective plus 1.6 × doubler by 3DHISTECH (Budapest, Hungary) PANNORAMIC® 1000 or 250 DX scanners and stored in MRXS format. The scanning resolution of WSIs was 0.25 µm / pixel.

An additional slide from 3 different tumors from the main cohort was cut and stained with HES in the Department of Pathology at Charles Nicolle Hospital in Rouen, France, to introduce another pre-analytical procedure.

Annotations and information extraction from WSIs

All WSIs were meticulously reviewed by 3 pathologists to exclude slides with noticeable artifacts or inadequate staining to ensure data quality. Faded slides were replaced by new sections from the original blocks.

Two experienced pathologists (C.G. and A.L.B.) and one junior pathologist (L. C.), blinded to the clinico-biological data, manually annotated all WSIs. Annotations were defined as follows, identifying three tissue classes: non-tumor liver (NT); tumor tissue with non-pejorative architecture (NP) i.e. non-macrotrabecular and non-VETC architecture, and tumor tissue with pejorative architecture (P) i.e. macrotrabecular and/or VETC architecture.

The macrotrabecular architecture is defined by hepatocyte trabeculae mostly being ≥10 cells thick according to the WHO Classification of Tumors: Digestive System Tumors 2019. VETC pattern is defined by clusters of tumor cells bordered by a complete rim of endothelial cells[15]. The two architectures can be combined in the same tumor[14].

To obtain a tumor architecture diagnostic model, the 107 patients were randomly divided into a training set (56%, i.e. 60 patients), a validation set (12%, i.e. 13 patients), and an independent test set (32%, i.e. 34 patients), with all WSIs from a given patient exclusively grouped within the same set. This data splitting method was employed to prevent data leakage between datasets.

The WSIs, stored in MRXS format, had an average size of $10^6 \times 10^6$ pixels. To extract information from training (n=399 WSIs) and validation (n=77 WSIs) sets, annotations were first made using fixed squares measuring 5888 × 5888 pixels (1400 µm × 1400 µm), with each rectangle containing

only one tissue class. The annotations were exported as TIFF format images, accompanied by their coordinates stored in a metadata file in XML format on an external encrypted hard drive.

The test set was annotated in two ways: with fixed squares as mentioned above for training and validation sets (n=81 WSIs) or with free-hand drawn polygons of different colors containing only one tissue class and covering the whole tissue section (n=123 WSIs) as shown in Figure 1. This latter annotation method made it possible to obtain metrics on the model performance at the level of entire slides and not just patches.

On slides annotated with fixed squares (557 WSIs from 87 patients), mVI images were also outlined by the 3 pathologists, allowing them to quantify the number of tumor emboli per tumor.

WSI processing (Figure 2)

Exported TIFF images of size 5888 × 5888 pixels cannot be processed quickly by standard machine learning algorithms. Each image was divided into non-overlapping square patches of 512 × 512 pixels also called tiles[22,23]. Non-overlapping square patches of the same dimension were derived directly from the whole surface of free-hand drawn polygon annotations.

The model analyzes images across multiple resolution: levels to reproduce the pathologist's observation at different scales[24]. Every single patch of 512 × 512 pixels was extended with an increased size determined by a growth factor $d = 1.5$. This operation was repeated to come up with three tiles at three distinct scales: 512 × 512 pixels, 768 × 768 pixels, and 1152 × 1152 pixels. Generated patches inherited their class from the annotated region. Finally, 768 × 768 pixels and 1152 × 1152 pixels patches were down-sampled to size 512 × 512 pixels by making use of an anti-aliasing filter[25,26], to generate a stack of three images of the same size.

After extracting patches from the images, a filtration step was implemented to eliminate patches that contained more than 80% white pixels. The rationale behind this filtration is to ensure that the analysis focuses on patches that contain relevant data, enhancing the overall reliability of the results.

The final step was an image data augmentation process[27–29], a standard technique allowing to enrich the dataset during training. The primary objective of data augmentation is to generate modified images that introduce new variations to the diagnostic model at each training iteration. To accomplish this, various spatial transformations, such as random vertical and horizontal flips, as well as random rotations with reflection mode within a rotation degree range of [-45, 45][30–33] were applied. A standardization method to scale down the input values in the data set to a nominal range between zero to one was used. This method is crucial not just due to potential variations in the features present across images, which could influence the model's outcomes[34], but also because it aids in maintaining the weights near zero, thereby enhancing the stability of the network during backpropagation[35].

Dataset description of the diagnostic model based on deep learning.

Characteristics of the 3 disjoint data sets: training set (60 patients), validation set (13 patients), and test set (34 patients) are described in Supplementary Information: Table S1. The percentages of non-tumor patches, non-pejorative architecture patches, and pejorative architecture patches in the training test were respectively 39.84%, 44.46%, and 15.70%.

Due to the class imbalance in the training set, the image quantity of the underrepresented classes in the training set was adjusted by combining two main techniques: oversampling the underrepresented categories by randomly duplicating non-tumor patches and pejorative patches[36] and using a weighted cross-entropy loss[37] defined as:

$$\boldsymbol{Loss}(\boldsymbol{y},\boldsymbol{p}) = \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}} \sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{C}} \boldsymbol{W_j} \left(\boldsymbol{y^{ij}}\, \boldsymbol{log}\left(\boldsymbol{p^{ij}}\right) + \left(\mathbf{1} - \boldsymbol{y^{ij}}\right) \boldsymbol{log}\left(\mathbf{1} - \boldsymbol{p^{ij}}\right)\right)$$

where $\boldsymbol{N}$ stands for the number of samples, $\boldsymbol{C}$ represents the number of classes (i.e., 3 in our study). $\boldsymbol{W_j}$ is the weight of the class $\boldsymbol{j}$, $\boldsymbol{y^{ij}}$ is the ground truth label of the sample $\boldsymbol{i}$ for the class $\boldsymbol{j}$, $\boldsymbol{p^{ij}}$ is the predicted probability of the sample $\boldsymbol{i}$ belonging to the class $\boldsymbol{j}$.

Neural network architecture (Figure 3)

The model input was structured around three parallel patches, each measuring 512 × 512 pixels. These patches were derived from the extended patches described in the WSI processing, distinguished by varying image magnification factors. The neural network was based on the ResNet architecture[38]. For the needs of this work, a ResNet34 was trained from scratch on the dataset. The architecture of this model involves three ResNet34 networks operating in parallel, where each input patch undergoes ResNet34 processing to generate a feature vector. The three generated feature vectors were then concatenated into a global vector. Subsequently, this global vector passed through three trainable fully connected layers, incorporating linear layers, a batch normalization layer, and a rectified linear unit (RELU) activation function[39]. Through this intricate process, the model produced the final output corresponding to the input class.

Ensembling Approach

The final stage consisted of an ensembling of the core ResNet34 (Supplementary Information: Figure S1). First, the training images and the validation images were merged and then split into 5-folds, out of which four were exclusively used for training, while the fifth served as a validation set. Importantly, the 5-folds split closely adhered to the 5-fold cross-validation methodology[40–42], ensuring a balanced representation of the three classes within each training fold.

For each pairing of the training and validation folds, each model was trained for 50 epochs with Adam optimizer[43]. The initial learning rate was set at 0.01 with a decrease of 0.1 when there was no improvement after 10 epochs. For neural network calibration purposes[44,45], a label smoothing technique[46–48] was used in the loss function with a smoothing parameter $\alpha = 0.125$.

For each training fold, weights were exclusively saved for the best-performing model, determined by the minimum value of the loss function. In this way, the weight for 5 folds corresponding to the five best models was saved. Afterward, a final prediction was generated as a probability score. To obtain the average probability, the mean of the outputs from the five final fully connected layers, also called logits, was computed. The SoftMax activation function[39] was finally applied to these mean values to obtain the probabilities indicating the likelihood of belonging to one of the three

predefined classes. Three types of means arithmetic, geometric, and harmonic were tested. Their formulas are described below:

Arithmetic mean: $\mu_{arithmetic} = \frac{1}{M}\sum_{i=1}^{M} p_i$

Geometric mean: $\mu_{geometric} = (\prod_{i=1}^{M} p_i)^{\frac{1}{M}}$

Harmonic mean: $\mu_{harmonic} = \frac{M}{\sum_{i=1}^{M}\frac{1}{p_i}}$

where $\boldsymbol{p_i}$ is the probability vector generated by model $\boldsymbol{i}$ and $\boldsymbol{M}$ is the number of models ($\boldsymbol{M = 5}$). The arithmetic operations were performed component-wise.

At inference, the five models were applied in parallel, and the type of mean giving the best metrics was used to classify all data.

Visualization of the model prediction on WSIs

A simplified visual heatmap of the WSI was generated to construct an interpretable and conclusive tumor segmentation mask for each WSI. During the initial tiling preprocessing step, each extracted patch's coordinates (x, y) were meticulously saved. Subsequently, the WSI was reconstructed using these stored (x, y) coordinates. In this process, every patch contributed to the final visualization, with each pixel on the heatmap assigned to a color corresponding to its predictive classification. The non-tumor liver tissue is represented in green, the non-pejorative tumor area in yellow, and the pejorative tumor area in red. This method provides a concise and intuitive display of the model predictions across the entirety of the WSI for the pathologists since it generates an interpretable tumor mask for each WSI.

Features extracted using the diagnostic model

From the heatmaps generated with the diagnostic model, two features were extracted as described below: the largest connected area of the pejorative architecture from each WSI of a given patient (AreaP) and the largest connected area of the pejorative architecture among all the WSIs of a given patient (AreaP_max).

These features were correlated with the presence of mVI and the exact number of tumor emboli when this information was available.

External validation cohort

Paraffine blocks from the external validation cohort of Henri-Mondor Hospital were cut and stained at Henri-Mondor. All slides (n=32) were digitized at 20 × objective by Hamamatsu (Hamamatsu, Japan) NanoZoomer S360 Digital slide scanner and stored in NDPI format. The scanning resolution of WSIs was 0.46 µm / pixel.

Those 32 WSIs from the external validation cohort were annotated using the second method of large free-hand drawn polygons.

Due to the variability introduced by a different pre-analytical procedure and a different scanner, the model trained with Paul-Brousse cohort could not be directly applied to Henri-Mondor cohort with satisfactory results. The primary difference between these cohorts lies in the scanner resolution. Consequently, to surmount the resolution disparity and facilitate the application of the model, a mathematical procedure known as interpolation was employed to mitigate the divergence in resolution by generating new pixels within the patches obtained from Henri-Mondor and thereby fostering a smoother transition during the upscaling of these patches to align with the resolution of the first cohort. In this instance, the spline interpolation method[49–52] was selected as the preferred mathematical approach to achieve the necessary resolution harmonization.

Evaluation metrics

For the model evaluation, confusion matrices[53,54] were used, facilitating a thorough examination of model performance by aligning its predictions with the actual ground truth labels and displaying the number of accurate and inaccurate instances. Plotting confusion matrices, rather than relying solely on accuracy, offers several advantages, especially with unbalanced data. This approach provides a more transparent and informative way to communicate model performance, highlighting areas of excellence and indicating areas for improvement. It offers a clear breakdown of true positives, true negatives, false positives, and false negatives for each class, providing a

comprehensive picture of the model performance. This matrix is crucial for computing essential classification metrics, including accuracy, balanced accuracy, F1 Score, precision, and sensitivity (Supplementary Information: Figure S2).

Results

Patient characteristics

Patients' baseline characteristics are summarized in Table 1.

For the Paul-Brousse cohort, among 107 patients, 87 (81%) were male. The underlying liver disease was predominantly viral hepatitis in 42% of cases followed by non-alcoholic fatty liver disease, and alcohol-related liver disease in 23% and 12% of cases respectively. In 11% of the patients, HCC occurred without an underlying liver disease. The non-tumor liver was F0-F2 in 36% and F3-F4 in 64% of cases according to METAVIR fibrosis score. Most patients had a single tumor. Tumor size ranged from 0.8 to 23 cm with a median size of 4 cm. WHO histological grade was 1 in 16 tumors (15%), 2 in 79 tumors (74%), and 3 in 12 tumors (11%). Microvascular tumor invasion was identified in 58 tumors (54%).

For the external cohort, among 29 patients, 22 (75%) were male. The underlying liver disease was mostly unknown (35%), then predominantly viral hepatitis in 34% of cases followed by alcohol-related liver disease and non-alcoholic fatty liver disease in 21% and 7% of cases respectively. Non-tumor liver was F0-F2 in 62% and F3-F4 in 38% of cases. Most patients had a single tumor. Tumor size ranged from 1.2 to 19.5 cm with a median size of 4.5 cm. WHO histological grade was 1 in 2 tumors (7%), 2 in 22 tumors (76%), and 3 in 5 tumors (17%). Microvascular tumor invasion was identified in 9 tumors (31%).

Performances of the algorithm at the patch-level (test subset 1)

The performance metrics of the 5-fold cross-validation before ensembling at the patch-level on test subset 1 are shown in Table 2. They are described with harmonic mean ensembling in Figure 4. The performance metrics with the three types of mean are detailed in Supplementary

Information: Table S2. Stable and consistent performances were observed whatever the employed mean. The models under evaluation demonstrated an accuracy metric above 0.860 reflecting the model's proficiency in correctly classifying instances. The arithmetic mean registered an accuracy of 0.862, and both the geometric and harmonic means achieved values of 0.864. For balanced accuracy, a metric accounting for imbalanced class distribution, the harmonic mean exhibited superior performance, achieving a score of 0.851, surpassing the corresponding values of 0.843 and 0.845 for the arithmetic and geometric means, respectively. Moreover, the F1 score achieved a satisfying value exceeding 0.82. The F1 metric is the harmonic mean of precision and recall, offering a balanced assessment of the model ability to correctly identify positive instances while minimizing false positives and false negatives. Considering these observations, the forthcoming presentation of results will prioritize those obtained through the harmonic mean, given it demonstrated superior performance compared to the other two means. However, it is essential to acknowledge that the arithmetic and geometric means yielded results that are approximately comparable to each other.

Performances of the algorithm at the WSI-level (test subset 2)

The performance metrics of the 5-fold cross-validation, before ensembling, at the WSI-level on test subset 2 are shown in Table 3. Furthermore, the performance metrics of the model at the WSI level on test subset 2 with harmonic mean are described in Figure 4. The performance metrics of the model at the WSI-level on test subset 2 for the three types of mean are detailed in Supplementary Information: Table S3. Again, stable and consistent performance regardless of the type of mean was observed with a slight superiority for the harmonic mean. The accuracy metric achieved a score of 0.823 and a balanced accuracy of 0.819. The F1 Score attained 0.851. The precision and sensitivity scores were 0.928 and 0.823.

The slight decrease in performance between the two test subsets at the patch-level and WSI-level could be attributed to the fact that polygonal annotations may contain some noise, as accurately limiting and precisely defining the borders between the three architectural patterns is challenging.

To clearly visualize the model prediction, Figure 5 showcases an example of a WSI processed by the model. Overall, the majority of the heatmap regions were successfully predicted with a high confidence level. Furthermore, in cases where the model partially misclassified certain zones, it reflected a low confidence level, serving as an alert to pathologists about the necessity of a human check for these regions.

Impact of the staining procedure on performances of the algorithm at WSI-level

In the endeavor of assessing the robustness of the model against domain shifts such as staining differences in distinct sites, an evaluation of the model was conducted on 3 WSIs from the test set colored in two different hospitals: Bicêtre and Rouen. The comparative analysis of the heatmaps from one of these WSI, shown as an example in Supplementary Information: Figure S3, reveals the resilience of the model against staining variability. The performance metrics show noteworthy achievements for both staining sites and a marginal superiority arises when evaluating the model on the WSIs stained at Bicêtre Hospital, which aligns with the training data distribution.

Performance of the algorithm on the external validation cohort

In this study, the robustness and efficacy of the algorithm predictions were validated by incorporating an external dataset from Henri-Mondor Hospital. This dataset, sourced from another renowned medical institution, served as a critical external validation for the model, providing a diverse set of cases for which the fixation procedure, the staining process, tissue section thickness, and scanner parameters were different. WSIs were annotated using polygonal annotations as test subset 2.

The evaluation of the model on the external validation cohort yielded commendable performances, affirming its generalization capabilities beyond the first cohort dataset. The accuracy achieved 0.787. The balanced accuracy reached a notable value of 0.731, indicative of the model's ability to maintain effectiveness across different classes. Furthermore, the F1 Score attained 0.812. The precision of 0.850 underscores the model's accuracy in correctly identifying positive instances, while the sensitivity of 0.802 emphasizes its capacity to capture a significant proportion of actual

positive cases. These results collectively highlight the model's robust performance and its potential utility in diverse real-world scenarios. The outcomes of this validation process are presented in Figure 6.

Correlation of pejorative architecture and presence of mVI

The correlation coefficients between various AI features related to pejorative architecture i.e. the largest connected area of the pejorative architecture from each WSI of a given patient and the largest connected area of the pejorative architecture among all the WSIs of a given patient were computed with the presence of mVI and the number of emboli. The correlation matrix is displayed in Figure 7.

The adverse architectural features exhibited a positive correlation with both the presence of mVI and the number of emboli. Notably, the dimensions of the adverse area (AreaP) and the maximum adverse area across all WSIs from patients (AreaP_max) emerged as viable predictors for mVI and the number of emboli. Specifically, the correlation coefficient (ranging from -1 to +1) for AreaP and mVI was 0.342, while the correlation coefficient for AreaP_max and mVI reached 0.384. Similarly, as regards the number of emboli, the correlation coefficients for AreaP and AreaP_max were 0.536 and 0.566, respectively. These coefficients underscore the potential predictive utility of AreaP and AreaP_max, yet simultaneously signal the complexity of predicting mVI.

Discussion

This study made it possible to develop a deep learning model capable of classifying non-tumor areas, tumor areas with non-pejorative architecture, and tumor areas with pejorative architecture of HCC on digital histopathological slides from liver surgical specimens. The choice was made for a supervised approach on a large sampling for each tumor to take into account tumor heterogeneity which is a hallmark of HCC and to obtain fully explicable results and best performances. In addition, the model was trained and validated on small and unequivocal square annotations and then trained on larger images. The dataset was substantial, encompassing more

than 700.000 tiles thus allowing a robust approach to address challenges associated with overfitting, a common concern in deep learning, particularly when dealing with limited data[55]. The presence of 3 annotators added diversity and thus limited classification bias. The neural network chosen was ResNet since it was designed for image classification and has been shown to be a powerful classifier, adopted in various computer vision applications in medical fields[56–60]. The final stage of the algorithm consisted of an ensembling of the core ResNet. Indeed, ensembling methods have gained interest today since it was proven that aggregating multiple algorithms is successful in increasing the overall prediction accuracy in a wide range of fields, especially in healthcare[61–63]. As a result, the model achieved a patch-level accuracy of 0.864 and a WSI-level accuracy of 0.823.

The automatic detection of pejorative architectures in HCC through the AI algorithm allows precise quantification within the entire tumor making it already a potential aid to diagnosis in the general context of pathologists shortage. Moreover, it generates AI features relevant to prognostic predictions, such as the largest connected area of pejorative architecture across all slides from a patient, i.e. the most extensive area of pejorative architecture. The model demonstrated robust performance on slides stained in another Department of Pathology using a different procedure, as well as for an external validation cohort.

This work marks the initial step in developing an accurate predictive tool for HCC recurrence after liver resection. Indeed, the availability of systemic therapies with promising results in HCC makes it essential to accurately stratify patients according to their risk of recurrence after surgical resection to identify the best candidates for adjuvant systemic treatment.

Among the different elements of a predictive model for the risk of recurrence, histopathological criteria are of paramount importance in this prognostication as they reflect the biological aggressiveness of the tumor. One of the main features in HCC is the presence of mVI, visible only at the microscopical level and defined by tumor invasion of vessels lined by endothelial cells. However, the variability in sampling surgical specimens from one pathologist to another[64], the poor

reproducibility in assessing mVI, and the lack of grading of this lesion have led to great heterogeneity in evaluating this valuable histological feature in HCC[65].

More recently, macrotrabecular architecture[12,13] and VETC pattern[14,15] have been identified as other markers of tumor aggressiveness. Macrotrabecular architecture had initially been defined as having trabeculae >6 cells thick[66] but the definition retained in the WHO classification published in 2019 is trabeculae being ≥10 cells thick. This discrepancy in the definition of macrotrabecular architecture may result in slightly different prevalence figures. Nevertheless, macrotrabecular architecture in at least 20% of tumor area is observed in around 50% of HCC[12,15]. The MTM HCC characterized by a predominant macrotrabecular growth pattern represents from 7 to 15% of HCC and is correlated with satellite nodules, macrovascular and microvascular invasion[12,67]. Jeon et al.[68] suggest that ≥30% of macrotrabecular architecture could be used as the more appropriate cut-off for defining MTM HCC. Whatever the definition, MTM HCC is associated with poorer overall survival, a higher recurrence rate, and a worse recurrence-free survival after liver resection[12,68].

VETC pattern characterized by clusters of tumor cells bordered by a complete rim of endothelial cells is also a pejorative pattern correlated with the presence of macrotrabecular architecture and mVI, and associated with poorer overall survival, poorer disease free survival and early recurrence[15]. This pattern is observed in 39% of HCC with a cut-off ≥5% and in 19% of HCC with a cut-off ≥55% which defines VETC phenotype. In addition, VETC pattern may act as a predictor of sorafenib benefit for HCC[69].

In a recent study, Chen et al.[70] developed a deep learning model to predict mVI in HCC from tumor areas of WSIs. Visualization results showed that macrotrabecular architecture with rich blood sinuses was one of the key features associated with mVI. Accordingly, a positive correlation between the presence of pejorative tumor architecture and the mVI status for the 107 patients of Paul-Brousse cohort was shown. All these data suggest that the identification of these two pejorative architectures within a tumor could have a strong predictive value for the risk of recurrence and could be an efficient surrogate of mVI. It is evident that achieving accurate

predictions necessitates more sophisticated features, such as nucleus features, beyond the scope of conventional AI features alone.

The digital transition of pathology has paved the way for the application of AI in pathology. Recent advances in machine learning, especially in deep neural networks, have enabled the identification of histopathological patterns through computer vision. Several recent studies have applied AI in the field of liver tumors. Part of them are segmentation studies aiming to distinguish HCC from adjacent non-tumor liver[71–77]. The dataset for these studies ranges from 50 to 1733 WSIs. Although it is difficult to compare these different studies because of the use of different metrics, their performances are good or excellent with an accuracy ranging from 0.88 to 0.97[16]. Another part of the AI studies in the field of HCC proposes algorithms predicting either recurrence risk after surgery[78–80] or patient survival[81]. These studies include a first step of tumor/non-tumor segmentation and a second step of weakly supervised model labeling recurrence/non-recurrence or survival/non-survival at the patient level with usually only one digital slide per patient.

In the current study, to develop a predictive tool for HCC recurrence after resection, the first step was to build an AI model able to identify 3 tissue classes on the WSIs from surgical specimens of HCC resection: non-tumor liver, non-pejorative tumor architecture, and pejorative tumor architecture including macrotrabecular and VETC patterns. Access to the architectural map will enable us to better understand the results returned by the algorithm and avoid the black-box effect. To the best of our knowledge to date, no AI algorithm has been published to classify the histological patterns of HCC. For other tumors such as lung cancer or glioblastoma, algorithms for classifying histological subtypes have been published with various levels of performance[82–85]. In the field of liver pathology, only three studies aimed to produce classification algorithms available to differentiate benign, dysplastic, and malignant hepatocellular nodules[18] and HCC from cholangiocarcinoma[17,86]. A recently published study[87] aimed at developing a deep pathomics score for predicting HCC recurrence after liver transplantation included a first step of classification

of tissue in 6 categories: normal liver tissue, portal area, fibrous tissue, immune cells, tumor region and hemorrhage/necrotic tissue but did not identify different tumor classes.

The approach of the current study is a fully supervised learning method based on over a hundred patients with a mean number of WSIs per patient over 6 whereas most studies included a single WSI per patient. The choice of this approach is based on the impossibility of labeling histological features at slide-level due to the heterogeneity of HCC. The performance of the model is stable and consistent whatever the employed mean for the ensembling method. The harmonic mean from the ensembling method slightly outperforms the geometric and the arithmetic means with an accuracy of 0.864, a balanced accuracy of 0.851, a F1 Score of 0.829, a precision of 0.814 and a sensitivity of 0.851. The confusion matrix indicates that the model is excellent to segment tumor versus non-tumor areas. The results are quite correct to differentiate tumor pejorative and non-pejorative architectures with 75% of pejorative areas being correctly identified. A possible explanation for the difficulty in distinguishing the two pejorative architectures is that one of the important visual criteria to identify them is the accentuated visibility of the sinusoidal lumens. This aspect may be obscured due to tissue compression phenomena or fixation artifacts and consequently missed during annotations.

Interestingly, the annotation method of the test set influenced the performance of the model. Test subset 1 was annotated as the training and validation sets through fixed squares of 5888 × 5888 pixels containing only one tissue class and including 25 patches of 512 × 512 pixels. Test subset 2 was annotated through large free-hand drawn polygonal areas containing only one tissue class and covering the whole tumor surface. This second annotation method allows to calculate the accuracy of the model at WSI-level. However, with this annotation method, the area borders between pejorative and non-pejorative tumor architectures cannot be very precisely defined, leading to a suboptimal ground truth and a slightly decreased but still high accuracy of 0.823.

External validation of AI algorithms is critical to ensure their robustness and assess their generalization capability. Digital histopathological slides may differ drastically from one

Department of Pathology to another due to pre-analytic process including fixation, temperature, section thickness, staining procedure, and also to digitization step according to the slide scanner type that governs the WSI format and resolution. The first step was to test the impact of a different staining process applied to the same slides. Tissue sections from 3 HCC of Paul-Brousse cohort were cut in the Department of Pathology at Bicêtre Hospital and stained in the Department of Pathology at Charles Nicolle Hospital in Rouen without affecting the model performances. In a second step, the model was validated on a completely independent series of HCC samples from Henri-Mondor Hospital. HES slides from 29 surgical liver specimens with HCC harvested at the same period as the surgical specimens of Paul-Brousse cohort were digitized with a Hamamatsu scanner providing a lesser resolution of 0.46 µm / pixel versus 0.25 µm / pixel. Due to the variability introduced by slide preparation and scanning, the model trained with Paul-Brousse cohort could not be directly applied to Henri-Mondor cohort with satisfactory results. The patches of 512 × 512 pixels were resized to take into account the difference in image resolution between the two cohorts. After these corrections, the model performances for the external cohort were quite satisfactory with an accuracy of 0.787 at WSI-level. The external validation through a real life external cohort is probably better than using The Cancer Genome Atlas collections which can introduce some biased behavior according to the utilized Deep Neural Networks with WSI classification based on their acquisition site[88]. In addition, the external cohort provides information on post-surgical recurrence and not only on patient survival as The Cancer Genome Atlas.

In conclusion, this work provides a practical and immediately applicable tool for pathologists, leveraging a highly supervised approach with extensive annotations provided by three expert pathologists, on multiple slides per case to account for tumor heterogeneity. The novelty of this work lies in its direct clinical relevance, enabling precise characterization and quantification of adverse architectures, a task previously unexplored. These characteristics, often challenging to discern with the naked eye, especially across multiple slides, are crucial for systematic and accurate recording in patient reports. Consequently, this algorithm facilitates the diagnosis of the

most aggressive histological subtypes of HCC with high precision and holds promise for developing a combined predictive score of early recurrence.

In terms of machine learning, the innovation stems from the smart integration of several techniques, which proved essential for achieving robust results and accommodating heterogeneous data from three different hospitals. The main point is a novel pipeline operating on patches. This includes a neural network architecture operating concurrently at multiple scales (three resolution levels), with results further enhanced through a sophisticated ensembling approach. Additionally, our methodology carefully addresses class imbalance in the data.

The following steps are to implement this algorithm in the routine workflow of a fully digitalized Department of Pathology and to create a predictive model of post-surgical recurrence including AI features from the algorithm, nuclear features obtained by image analysis, macroscopic tumor characteristics, and biological data.

Conflict of Interest

The authors declare no funding or conflicts of interest.

Author Contributions

A.L.B., A.S., C.G., and J.C.P. performed study concept and design and wrote the first draft of the article; A.L.B., L.C. and C.G. provided acquisition of data; A.S. and J.C.P. provided analysis, interpretation of data, and statistical analysis; L.C., O.R. and M.L. participated in the study design; E.V. and D.C. performed review of the paper; K.P. provided technical support; J.C. provided the external validation cohort.

C.G. is the guarantor of this work and, as such, had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

All authors read and approved the final paper.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F: Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021, 71:209–249.

2. Zhang X, El-Serag HB, Thrift AP: Predictors of five-year survival among patients with hepatocellular carcinoma in the United States: an analysis of SEER-Medicare. Cancer Causes Control 2021, 32:317–325.

3. Gelli M, Sebagh M, Porcher R, Romanelli E, Vibert E, Sa Cunha A, Castaing D, Rosmorduc O, Samuel D, Adam R, Cherqui D: Liver Resection for Early Hepatocellular Carcinoma: Preoperative Predictors of Non Transplantable Recurrence and Implications for Treatment Allocation. Ann Surg 2020, 272:820–826.

4. Llovet JM, Kelley RK, Villanueva A, Singal AG, Pikarsky E, Roayaie S, Lencioni R, Koike K, Zucman-Rossi J, Finn RS: Hepatocellular carcinoma. Nat Rev Dis Primers 2021, 7:6.

5. Roayaie S, Obeidat K, Sposito C, Mariani L, Bhoori S, Pellegrinelli A, Labow D, Llovet JM, Schwartz M, Mazzaferro V: Resection of hepatocellular cancer ≤2 cm: results from two Western centers. Hepatology 2013, 57:1426–1435.

6. Finn RS, Qin S, Ikeda M, Galle PR, Ducreux M, Kim T-Y, Kudo M, Breder V, Merle P, Kaseb AO, Li D, Verret W, Xu D-Z, Hernandez S, Liu J, Huang C, Mulla S, Wang Y, Lim HY, Zhu AX, Cheng A-L, IMbrave150 Investigators: Atezolizumab plus Bevacizumab in Unresectable Hepatocellular Carcinoma. N Engl J Med 2020, 382:1894–1905.

7. Casadei-Gardini A, Rimini M, Tada T, Suda G, Shimose S, Kudo M, Cheon J, Finkelmeier F, Lim HY, Rimassa L, Presa J, Masi G, Yoo C, Lonardi S, Tovoli F, Kumada T, Sakamoto N, Iwamoto H, Aoki T,

Chon HJ, Himmelsbach V, Pressiani T, Montes M, Vivaldi C, Soldà C, Piscaglia F, Hiraoka A, Sho T, Niizeki T, Nishida N, Steup C, Iavarone M, Di Costanzo G, Marra F, Scartozzi M, Tamburini E, Cabibbo G, Foschi FG, Silletta M, Hirooka M, Kariyama K, Tani J, Atsukawa M, Takaguchi K, Itobayashi E, Fukunishi S, Tsuji K, Ishikawa T, Tajiri K, Ochi H, Yasuda S, Toyoda H, Ogawa C, Nishimura T, Hatanaka T, Kakizaki S, Shimada N, Kawata K, Tada F, Ohama H, Nouso K, Morishita A, Tsutsui A, Nagano T, Itokawa N, Okubo T, Arai T, Imai M, Kosaka H, Naganuma A, Koizumi Y, Nakamura S, Kaibori M, Iijima H, Hiasa Y, Burgio V, Persano M, Della Corte A, Ratti F, De Cobelli F, Aldrighetti L, Cascinu S, Cucchetti A: Atezolizumab plus bevacizumab versus lenvatinib for unresectable hepatocellular carcinoma: a large real-life worldwide population. Eur J Cancer 2023, 180:9–20.

8. Qin S, Chen M, Cheng A-L, Kaseb AO, Kudo M, Lee HC, Yopp AC, Zhou J, Wang L, Wen X, Heo J, Tak WY, Nakamura S, Numata K, Uguen T, Hsiehchen D, Cha E, Hack SP, Lian Q, Ma N, Spahn JH, Wang Y, Wu C, Chow PKH, IMbrave050 investigators: Atezolizumab plus bevacizumab versus active surveillance in patients with resected or ablated high-risk hepatocellular carcinoma (IMbrave050): a randomised, open-label, multicentre, phase 3 trial. Lancet 2023, 402:1835–1847.

9. Xu X, Zhang H-L, Liu Q-P, Sun S-W, Zhang J, Zhu F-P, Yang G, Yan X, Zhang Y-D, Liu X-S: Radiomic analysis of contrast-enhanced CT predicts microvascular invasion and outcome in hepatocellular carcinoma. J Hepatol 2019, 70:1133–1144.

10. Lewin M, Laurent-Bellue A, Desterke C, Radu A, Feghali JA, Farah J, Agostini H, Nault J-C, Vibert E, Guettier C: Evaluation of perfusion CT and dual-energy CT for predicting microvascular invasion of hepatocellular carcinoma. Abdom Radiol (NY) 2022, 47:2115–2127.

11. Erstad DJ, Tanabe KK: Prognostic and Therapeutic Implications of Microvascular Invasion in Hepatocellular Carcinoma. Ann Surg Oncol 2019, 26:1474–1493.

12. Calderaro J, Couchy G, Imbeaud S, Amaddeo G, Letouzé E, Blanc J-F, Laurent C, Hajji Y, Azoulay D, Bioulac-Sage P, Nault J-C, Zucman-Rossi J: Histological subtypes of hepatocellular carcinoma are related to gene mutations and molecular tumour classification. Journal of Hepatology 2017, 67:727–738.

13. Ziol M, Poté N, Amaddeo G, Laurent A, Nault J-C, Oberti F, Costentin C, Michalak S, Bouattour M, Francoz C, Pageaux GP, Ramos J, Decaens T, Luciani A, Guiu B, Vilgrain V, Aubé C, Derman J, Charpy C, Zucman-Rossi J, Barget N, Seror O, Ganne-Carrié N, Paradis V, Calderaro J: Macrotrabecular-massive hepatocellular carcinoma: A distinctive histological subtype with clinical relevance. Hepatology 2018, 68:103–112.

14. Fang J-H, Zhou H-C, Zhang C, Shang L-R, Zhang L, Xu J, Zheng L, Yuan Y, Guo R-P, Jia W-H, Yun J-P, Chen M-S, Zhang Y, Zhuang S-M: A novel vascular pattern promotes metastasis of hepatocellular carcinoma in an epithelial-mesenchymal transition-independent manner. Hepatology 2015, 62:452–465.

15. Renne SL, Woo HY, Allegra S, Rudini N, Yano H, Donadon M, Viganò L, Akiba J, Lee HS, Rhee H, Park YN, Roncalli M, Di Tommaso L: Vessels Encapsulating Tumor Clusters (VETC) Is a Powerful Predictor of Aggressive Hepatocellular Carcinoma. Hepatology 2020, 71:183–195.

16. Allaume P, Rabilloud N, Turlin B, Bardou-Jacquet E, Loréal O, Calderaro J, Khene Z-E, Acosta O, De Crevoisier R, Rioux-Leclercq N, Pecot T, Kammerer-Jacquet S-F: Artificial Intelligence-Based Opportunities in Liver Pathology-A Systematic Review. Diagnostics (Basel) 2023, 13:1799.

17. Kiani A, Uyumazturk B, Rajpurkar P, Wang A, Gao R, Jones E, Yu Y, Langlotz CP, Ball RL, Montine TJ, Martin BA, Berry GJ, Ozawa MG, Hazard FK, Brown RA, Chen SB, Wood M, Allard LS, Ylagan L, Ng AY, Shen J: Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPJ Digit Med 2020, 3:23.

18. Cheng N, Ren Y, Zhou J, Zhang Y, Wang D, Zhang X, Chen B, Liu F, Lv J, Cao Q, Chen S, Du H, Hui D, Weng Z, Liang Q, Su B, Tang L, Han L, Chen J, Shao C: Deep Learning-Based Classification of Hepatocellular Nodular Lesions on Whole-Slide Histopathologic Images. Gastroenterology 2022, 162:1948-1961.e7.

19. Organisation mondiale de la santé, Centre international de recherche sur le cancer, editors: Digestive system tumours. 5th ed. Lyon, International agency for research on cancer, 2019, .

20. Amin MB, American Joint Committee on Cancer, American Cancer Society, editors: AJCC cancer staging manual. Eight edition / editor-in-chief, Mahul B. Amin, MD, FCAP ; editors, Stephen B. Edge, MD, FACS [and 16 others] ; Donna M. Gress, RHIT, CTR-Technical editor ; Laura R. Meyer, CAPM-Managing editor. Chicago IL, American Joint Committee on Cancer, Springer, 2017, .

21. Intraobserver and interobserver variations in liver biopsy interpretation in patients with chronic hepatitis C. The French METAVIR Cooperative Study Group. Hepatology 1994, 20:15–20.

22. Lee ALS, To CCK, Lee ALH, Li JJX, Chan RCK: Model architecture and tile size selection for convolutional neural network training for non-small cell lung cancer detection on whole slide images. Informatics in Medicine Unlocked 2022, 28:100850.

23. Smith B, Hermsen M, Lesser E, Ravichandar D, Kremers W: Developing image analysis pipelines of whole-slide images: Pre- and post-processing. J Clin Trans Sci 2021, 5:e38.

24. Mnih V, Heess N, Graves A, kavukcuoglu koray: Recurrent Models of Visual Attention. Advances in Neural Information Processing Systems, Montréal, Canada, Curran Associates, Inc., 2014, pp. 2204–2212.

25. Ryan O: Applications of Antialiasing in an Image Processing Framework Setting. Proceedings of the 7th Nordic Signal Processing Symposium - NORSIG 2006, Reykjavik, Iceland, 2006, pp. 106–109.

26. Wu X: An efficient antialiasing technique. Acm Siggraph Computer Graphics, ACM New York, NY, USA, 1991, 25:143–152.

27. Wang J, Perez L: The effectiveness of data augmentation in image classification using deep learning. Convolutional Neural Networks Vis Recognit 2017, 11:1–8.

28. Shorten C, Khoshgoftaar TM: A survey on Image Data Augmentation for Deep Learning. J Big Data 2019, 6:60.

29. Mikolajczyk A, Grochowski M: Data augmentation for improving deep learning in image classification problem. 2018 International Interdisciplinary PhD Workshop (IIPhDW), Swinoujście, IEEE, 2018, pp. 117–122.

30. Toma TA, Biswas S, Miah MS, Alibakhshikenari M, Virdee BS, Fernando S: Breast Cancer Detection Based on Simplified Deep Learning Technique With Histopathological Image Using BreaKHis Database. Radio Science 2023, 58:e2023RS007761.

31. Gupta V, Bhavsar A: Sequential Modeling of Deep Features for Breast Cancer Histopathological Image Classification. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, IEEE, 2018, pp. 2335–23357.

32. Bardou D, Zhang K, Ahmad SM: Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks. IEEE Access 2018, 6:24680–24693.

33. Ameen YA, Badary DM, Abonnoor AEI, Hussain KF, Sewisy AA: Which data subset should be augmented for deep learning? a simulation study using urothelial cell carcinoma histopathology images. BMC Bioinformatics 2023, 24:75.

34. Yousif MJ: Enhancing The Accuracy of Image Classification Using Deep Learning and Preprocessing Methods. AIRDJ 2024, 3.

35. Huang L, Qin J, Zhou Y, Zhu F, Liu L, Shao L: Normalization Techniques in Training DNNs: Methodology, Analysis and Application. IEEE Trans Pattern Anal Mach Intell 2023, 45:10173–10196.

36. Mohammed R, Rawashdeh J, Abdullah M: Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, IEEE, 2020, pp. 243–248.

37. Ho Y, Wookey S: The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling. IEEE Access 2020, 8:4806–4813.

38. He K, Zhang X, Ren S, Sun J: Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, IEEE, 2016, pp. 770–778.

39. Nwankpa CE, Gachagan A, Marshall S: Activation Functions: Comparison of Trends in Practice and Research for Deep Learning. 2018, abs/1811.03378:124–133.

40. Berrar D: Cross-Validation. Encyclopedia of Bioinformatics and Computational Biology, Elsevier, 2019, pp. 542–545.

41. Kohavi R: A study of cross-validation and bootstrap for accuracy estimation and model selection. 105555/16430311643047, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1145.

42. Refaeilzadeh P, Tang L, Liu H: Cross-Validation. In: LIU L, ÖZSU MT, editors. Encyclopedia of Database Systems, Boston, MA, Springer US, 2009, pp. 532–538.

43. Kingma DP, Ba J: Adam: A Method for Stochastic Optimization. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, arXiv, 2015, .

44. Guo C, Pleiss G, Sun Y, Weinberger KQ: On Calibration of Modern Neural Networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 2017, pp. 1321–1330.

45. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z: Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, IEEE, 2016, pp. 2818–2826.

46. Liu B, Ayed IB, Galdran A, Dolz J: The Devil is in the Margin: Margin-based Label Smoothing for Network Calibration. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, IEEE, 2022, pp. 80–88.

47. Minderer M, Djolonga J, Romijnders R, Hubis F, Zhai X, Houlsby N, Tran D, Lucic M: Revisiting the Calibration of Modern Neural Networks. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW, editors. Advances in Neural Information Processing Systems, Curran Associates, Inc., 2021, pp. 15682–15694.

48. Müller R, Kornblith S, Hinton GE: When does label smoothing help? In: Wallach H, Larochelle H, Beygelzimer A, Alché-Buc F d', Fox E, Garnett R, editors. Advances in Neural Information Processing Systems, Vancouver, Canada, Curran Associates, Inc., 2019, pp. 4694–4703.

49. Costantini P: On Monotone and Convex Spline Interpolation. Mathematics of Computation, American Mathematical Society, 1986, 46:203–214.

50. Habermann C, Kindermann F: Multidimensional Spline Interpolation: Theory and Applications. Comput Econ 2007, 30:153–169.

51. Briand T, Monasse P: Theory and Practice of Image B-Spline Interpolation. Image Processing On Line 2018, 8:99–141.

52. Lung-Jen Wang, Wen-Shyong Hsieh, Trieu-Kien Truong, I. S. Reed, T. C. Cheng: A fast efficient computation of cubic-spline interpolation in image codec. IEEE Transactions on Signal Processing 2001, 49:1189–1197.

53. Stehman SV: Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment 1997, 62:77–89.

54. Powers D: Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation (Tech. Rep.). Adelaide, Australia 2007, .

55. Li H, Rajbahadur GK, Lin D, Bezemer C-P, Jiang ZM: Keeping Deep Learning Models in Check: A History-Based Approach to Mitigate Overfitting. IEEE Access 2024, :1–1.

56. Abousamra S, Gupta R, Hou L, Batiste R, Zhao T, Shankar A, Rao A, Chen C, Samaras D, Kurc T, Saltz J: Deep Learning-Based Mapping of Tumor Infiltrating Lymphocytes in Whole Slide Images of 23 Types of Cancer. Front Oncol 2022, 11:806603.

57. Petríková D, Cimrák I: Survey of Recent Deep Neural Networks with Strong Annotated Supervision in Histopathology. Computation 2023, 11:81.

58. Panigrahi S, Bhuyan R, Kumar K, Nayak J, Swarnkar T: Multistage classification of oral histopathological images using improved residual network. MBE 2021, 19:1909–1925.

59. Wang Z, Gao J, Kan H, Huang Y, Tang F, Li W, Yang F: ResNet for Histopathologic Cancer Detection, the Deeper, the Better? JDSIS 2023, .

60. Bidart R, Wong A: TriResNet: A Deep Triple-Stream Residual Network for Histopathology Grading. 2019, pp. 369–382.

61. Nahar N, Ara F, Neloy MdAI, Barua V, Hossain MS, Andersson K: A Comparative Analysis of the Ensemble Method for Liver Disease Prediction. 2019 2nd International Conference on Innovation in Engineering and Technology (ICIET), Dhaka, Bangladesh, IEEE, 2019, pp. 1–6.

62. Das A, Narayan Mohanty M, Kumar Mallick P, Tiwari P, Muhammad K, Zhu H: Breast cancer detection using an ensemble deep learning method. Biomedical Signal Processing and Control 2021, 70:103009.

63. Albashish D: Ensemble of adapted convolutional neural networks (CNN) methods for classifying colon histopathological images. PeerJ Computer Science 2022, 8:e1031.

64. Hu H-T, Wang Z, Kuang M, Wang W: Need for normalization: the non-standard reference standard for microvascular invasion diagnosis in hepatocellular carcinoma. World J Surg Oncol 2018, 16:50.

65. Rodríguez-Perálvarez M, Luong TV, Andreana L, Meyer T, Dhillon AP, Burroughs AK: A systematic review of microvascular invasion in hepatocellular carcinoma: diagnostic and prognostic variability. Ann Surg Oncol 2013, 20:325–339.

66. Kishi K, Shikata T, Hirohashi S, Hasegawa H, Yamazaki S, Makuuchi M: Hepatocellular carcinoma. A clinical and pathologic analysis of 57 hepatectomy cases. Cancer 1983, 51:542–548.

67. Sessa A, Mulé S, Brustia R, Regnault H, Galletto Pregliasco A, Rhaiem R, Leroy V, Sommacale D, Luciani A, Calderaro J, Amaddeo G: Macrotrabecular-Massive Hepatocellular Carcinoma: Light and Shadow in Current Knowledge. J Hepatocell Carcinoma 2022, 9:661–670.

68. Jeon Y, Benedict M, Taddei T, Jain D, Zhang X: Macrotrabecular Hepatocellular Carcinoma: An Aggressive Subtype of Hepatocellular Carcinoma. Am J Surg Pathol 2019, 43:943–948.

69. Fang J-H, Xu L, Shang L-R, Pan C-Z, Ding J, Tang Y-Q, Liu H, Liu C-X, Zheng J-L, Zhang Y-J, Zhou Z-G, Xu J, Zheng L, Chen M-S, Zhuang S-M: Vessels That Encapsulate Tumor Clusters (VETC) Pattern Is a Predictor of Sorafenib Benefit in Patients with Hepatocellular Carcinoma. Hepatology 2019, 70:824–839.

70. Chen Q, Xiao H, Gu Y, Weng Z, Wei L, Li B, Liao B, Li J, Lin J, Hei M, Peng S, Wang W, Kuang M, Chen S: Deep learning for evaluation of microvascular invasion in hepatocellular carcinoma from tumor areas of histology images. Hepatol Int 2022, 16:590–602.

71. Feng Y, Hafiane A, Laurent H: A deep learning based multiscale approach to segment the areas of interest in whole slide images. Comput Med Imaging Graph 2021, 90:101923.

72. Roy M, Kong J, Kashyap S, Pastore VP, Wang F, Wong KCL, Mukherjee V: Convolutional autoencoder based model HistoCAE for segmentation of viable tumor regions in liver whole-slide images. Sci Rep 2021, 11:139.

73. Wang X, Fang Y, Yang S, Zhu D, Wang M, Zhang J, Tong K-Y, Han X: A hybrid network for automatic hepatocellular carcinoma segmentation in H&E-stained whole slide images. Med Image Anal 2021, 68:101914.

74. Liao H, Xiong T, Peng J, Xu L, Liao M, Zhang Z, Wu Z, Yuan K, Zeng Y: Classification and Prognosis Prediction from Histopathological Images of Hepatocellular Carcinoma by a Fully Automated Pipeline Based on Machine Learning. Ann Surg Oncol 2020, 27:2359–2369.

75. Yang T-L, Tsai H-W, Huang W-C, Lin J-C, Liao J-B, Chow N-H, Chung P-C: Pathologic liver tumor detection using feature aligned multi-scale convolutional network. Artif Intell Med 2022, 125:102244.

76. Diao S, Tian Y, Hu W, Hou J, Lambo R, Zhang Z, Xie Y, Nie X, Zhang F, Racoceanu D, Qin W: Weakly Supervised Framework for Cancer Region Detection of Hepatocellular Carcinoma in Whole-Slide Pathologic Images Based on Multiscale Attention Convolutional Neural Network. Am J Pathol 2022, 192:553–563.

77. Feng S, Yu X, Liang W, Li X, Zhong W, Hu W, Zhang H, Feng Z, Song M, Zhang J, Zhang X: Development of a Deep Learning Model to Assist With Diagnosis of Hepatocellular Carcinoma. Front Oncol 2021, 11:762733.

78. Shi J-Y, Wang X, Ding G-Y, Dong Z, Han J, Guan Z, Ma L-J, Zheng Y, Zhang L, Yu G-Z, Wang X-Y, Ding Z-B, Ke A-W, Yang H, Wang L, Ai L, Cao Y, Zhou J, Fan J, Liu X, Gao Q: Exploring prognostic indicators in the pathological images of hepatocellular carcinoma based on deep learning. Gut 2021, 70:951–961.

79. Lu L, Daigle BJ: Prognostic analysis of histopathological images using pre-trained convolutional neural networks: application to hepatocellular carcinoma. PeerJ 2020, 8:e8668.

80. Yamashita R, Long J, Saleem A, Rubin DL, Shen J: Deep learning predicts postsurgical recurrence of hepatocellular carcinoma from digital histopathologic images. Sci Rep 2021, 11:2047.

81. Saillard C, Schmauch B, Laifa O, Moarii M, Toldo S, Zaslavskiy M, Pronier E, Laurent A, Amaddeo G, Regnault H, Sommacale D, Ziol M, Pawlotsky J-M, Mulé S, Luciani A, Wainrib G, Clozel T, Courtiol P, Calderaro J: Predicting Survival After Hepatocellular Carcinoma Resection Using Deep Learning on Histological Slides. Hepatology 2020, 72:2000–2013.

82. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, Moreira AL, Razavian N, Tsirigos A: Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018, 24:1559–1567.

83. Wei JW, Tafe LJ, Linnik YA, Vaickus LJ, Tomita N, Hassanpour S: Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. Sci Rep 2019, 9:3358.

84. Jin L, Shi F, Chun Q, Chen H, Ma Y, Wu S, Hameed NUF, Mei C, Lu J, Zhang J, Aibaidula A, Shen D, Wu J: Artificial intelligence neuropathologist for glioma classification using deep learning on hematoxylin and eosin stained slide images and molecular markers. Neuro Oncol 2021, 23:44–52.

85. Lami K, Ota N, Yamaoka S, Bychkov A, Matsumoto K, Uegami W, Munkhdelger J, Seki K, Sukhbaatar O, Attanoos R, Berezowska S, Brcic L, Cavazza A, English JC, Fabro AT, Ishida K, Kashima Y, Kitamura Y, Larsen BT, Marchevsky AM, Miyazaki T, Morimoto S, Ozasa M, Roden AC, Schneider F, Smith ML, Tabata K, Takano AM, Tanaka T, Tsuchiya T, Nagayasu T, Sakanashi H, Fukuoka J: Standardized Classification of Lung Adenocarcinoma Subtypes and Improvement of Grading Assessment Through Deep Learning. Am J Pathol 2023, 193:2066–2079.

86. Beaufrère A, Ouzir N, Zafar PE, Laurent-Bellue A, Albuquerque M, Lubuela G, Grégory J, Guettier C, Mondet K, Pesquet J-C, Paradis V: Primary liver cancer classification from routine tumour biopsy using weakly supervised deep learning. JHEP Reports 2024, 6:101008.

87. Qu W-F, Tian M-X, Lu H-W, Zhou Y-F, Liu W-R, Tang Z, Yao Z, Huang R, Zhu G-Q, Jiang X-F, Tao C-Y, Fang Y, Gao J, Wu X-L, Chen J-F, Zhao Q-F, Yang R, Chu T-H, Zhou J, Fan J, Yu J-H, Shi Y-H: Development of a deep pathomics score for predicting hepatocellular carcinoma recurrence after liver transplantation. Hepatol Int 2023, 17:927–941.

88. Asilian Bidgoli A, Rahnamayan S, Dehkharghanian T, Grami A, Tizhoosh HR: Bias reduction in representation of histopathology images using deep feature selection. Sci Rep 2022, 12:19994.

Figure legends

Figure 1:

Annotation methods on a Whole Slide Image stained with hematoxylin, eosin, and saffron; non-tumor liver is in green, tumor tissue with non-pejorative architecture is in yellow, and tumor tissue with pejorative architecture is in red. Fixed square annotations containing only one tissue class at 5 × magnification (A), large free-hand drawn polygons containing only one tissue class and covering the whole tissue section at 5 × magnification (B) and examples of non-tumor liver (green square), tumor tissue with non-pejorative architecture (yellow square), tumor tissue with pejorative architecture (red square) at 200 × magnification (C).

Figure 2: Flowchart of the data collection and whole slide images (WSIs) processing; on tumor segmentation, non-tumor liver (NT) appears in green, tumor tissue with non-pejorative architecture (NP) in yellow and tumor tissue with pejorative architecture (P) in red.

Figure 3: Neural network based on the ResNet architecture (NT: Non-Tumor, NP: Non-Pejorative, P: Pejorative).

Figure 4: Performances (A) and confusion matrices after ensembling on test subset 1 i.e. patch-level (B) and test subset 2 i.e. Whole Slide Image (WSI)-level (C).

Figure 5: Example of the model predictions on a Whole Slide Image and its performances with non-tumor liver in green, tumor tissue with non-pejorative architecture in yellow, and tumor tissue with pejorative architecture in red: the Whole Slide Image and the ground truth (A), model predictions (B), model confidence level (C), model performances on this Whole Slide Image (D) and confusion matrix (E).

Figure 6: Performances (A) and confusion matrix (B) for Henri-Mondor Hospital Whole Slide Images.

Figure 7: Correlation matrix between microvascular invasion (mVI), number of emboli and artificial intelligence features of pejorative architecture i.e. the largest connected area of the pejorative architecture from each Whole Slide Image of a given patient (AreaP) and the largest connected

area of the pejorative architecture among all the Whole Slide Images of a given patient (AreaP_max).

Table 1: Clinico-biological and pathological data of the internal and external cohorts.

		Bicêtre	Henri-Mondor
Number of patients (men / women)		107 (87 / 20)	29 (22 / 7)
Mean age (min - max)		66 (21 - 86)	68 (40 - 85)
Etiology of the underlying liver disease	Alcohol-related liver disease (%)	13 (12)	6 (21)
	Non-alcoholic fatty liver disease (%)	25 (23)	2 (7)
	Hepatitis B (%)	20 (19)	5 (17)
	Hepatitis C (%)	25 (23)	5 (17)
	Mixed etiology (%)	6 (6)	1 (3)
	Other etiology (%)	6 (6)	10 (35)
	Normal liver (%)	12 (11)	0 (0)
Serum alpha-foetoprotein (µg / L) N < 7	Data available (%)	106 (99)	21 (72)
	Median value (min-max)	6.9 (1 - 98,000)	6.0 (1 - 60,000)
Albumin (g / L) N: 35-45	Data available (%)	101 (94)	16 (55)
	Mean value (min-max)	39.5 (24.6 - 49.9)	39.5 (31 - 51)
Total bilirubin (µmol / L) N < 17	Data available (%)	106 (99)	16 (55)
	Mean value (min-max)	12.0 (5 - 36)	9.4 (4 - 21)
Number of hepatocellular carcinoma nodules (mean per patient)		116 (1.1)	36 (1.2)
Median size in cm (min - max)		4 (0.8 - 23)	4.5 (1.2 - 19.5)

Satellite nodules (%)		39 (36)	8 (28)
Multinodular expansive macroscopic pattern (%)		46 (43)	7 (24)
Hepatocellular carcinoma histological grade[19]	1 (%)	16 (15)	2 (7)
	2 (%)	79 (74)	22 (76)
	3 (%)	12 (11)	5 (17)
Hepatocellular carcinoma subtype[19]	Steatohepatitic (%)	8 (7)	1 (3)
	Clear cell (%)	3 (3)	2 (7)
	Macrotrabecular massive (%)	9 (8)	3 (10)
	Lymphocyte-rich (%)	5 (5)	5 (17)
Microvascular invasion (%)		58 (54)	9 (31)
Hepatocellular carcinoma stage[20]	T1a (%)	12 (11)	4 (14)
	T1b (%)	39 (36)	13 (45)
	T2 (%)	46 (43)	10 (34)
	T3 (%)	3 (3)	2 (7)
	T4 (%)	7 (7)	0 (0)
Non-tumor liver parenchyma fibrosis score[21]	0-2 (%)	38 (36)	18 (62)
	3-4 (%)	69 (64)	11 (38)

Table 2: Performances of 5-Fold cross validation on test subtest 1 (patch-level).

	Accuracy	**Balanced Accuracy**	**F1 Score**	**Precision**	**Sensitivity**
Fold 1	0.822	0.814	0.798	0.787	0.813
Fold 2	0.833	0.794	0.785	0.778	0.795
Fold 3	0.788	0.760	0.740	0.728	0.761
Fold 4	0.822	0.817	0.788	0.774	0.818
Fold 5	0.841	0.840	0.798	0.789	0.841

Table 3: Performances of 5-Fold cross validation on test subtest 2 (WSI-level).

	Accuracy	**Balanced Accuracy**	**F1 Score**	**Precision**	**Sensitivity**
Fold 1	0.803	0.803	0.789	0.778	0.802
Fold 2	0.814	0.785	0.777	0.772	0.785
Fold 3	0.770	0.751	0.732	0.718	0.760
Fold 4	0.801	0.804	0.775	0.762	0.808
Fold 5	0.822	0.834	0.788	0.785	0.837